

# Principal component gene set enrichment (PCGSE)

H. Robert Frost <sup>\*†‡</sup>, Zhigang Li<sup>\*†</sup> and Jason H. Moore <sup>\*†‡</sup>

March 21, 2014

## Abstract

**Motivation:** Although principal component analysis (PCA) is widely used for the dimensional reduction of biomedical data, interpretation of PCA results remains daunting. Most existing methods for interpreting principal components (PCs) attempt to explain each PC in terms of a small number of variables by generating approximate PCs with few non-zero loadings. Although these methods are useful when just a few variables dominate the population PCs, they are often inadequate for characterizing the biological signal represented by the PCs of high-dimensional genomic data. For genomic data, reproducible and biologically meaningful PC interpretation requires methods based on the combined signal of functionally related sets of genes. While gene set testing methods have been widely used in supervised settings to quantify the association of groups of genes with clinical outcomes, these methods have seen only limited application for testing the enrichment of gene sets relative to sample PCs.

**Results:** We describe a novel approach, principal component gene set enrichment (PCGSE), for computing the statistical enrichment or depletion of gene sets relative to PCs computed from genomic data. The PCGSE method performs a two-stage competitive gene set test using the correlation between each gene and each PC as the gene-level test statistic with flexible choice of both the gene set test statistic and the method used to compute the null distribution of the gene set statistic. Using simulated data with simulated gene sets and real gene expression data with curated gene sets, we demonstrate that biologically meaningful and computationally efficient results can be obtained from a simple parametric version of the PCGSE method that performs a correlation-adjusted two-sample t-test between the gene-level test statistics for gene set members and genes not in the set.

**Availability:** <http://cran.r-project.org/web/packages/PCGSE/index.html>

**Contact:** rob.frost@dartmouth.edu or jason.h.moore@dartmouth.edu

## 1 Introduction

Developed independently by Karl Pearson (Pearson, 1901) and Harold Hotelling (Hotelling, 1933), PCA is a well established statistical technique that performs a linear transformation of multivariate data into a new set of variables, the principal components (PCs), that are linear combinations of the original variables, are uncorrelated and have sequentially maximum variance (Jolliffe, 2002). The solution to PCA is given by the spectral decomposition of the covariance matrix with the variance of the PCs specified by the eigenvalues, arranged in decreasing order, and the PC directions specified by the associated eigenvectors.

---

<sup>\*</sup>Institute for Quantitative Biomedical Sciences, Geisel School of Medicine, Lebanon, NH 03756

<sup>†</sup>Section of Biostatistics and Epidemiology, Department of Community and Family Medicine, Geisel School of Medicine, Lebanon, NH 03756

<sup>‡</sup>Department of Genetics, Dartmouth College, Hanover, NH 03755

In the biomedical domain, PCA has been extensively employed for the analysis of genomic data including measures of DNA variation, DNA methylation, RNA expression and protein abundance (Ma and Dai, 2011). Common features of these datasets, and the motivation for spectral decomposition methods, are the high dimensionality of the feature space (i.e., from thousands to over one million), comparatively low sample size (i.e.,  $p \gg n$ ) and significant collinearity between the features. The most common uses of PCA with genomic data involve dimensionality reduction for visualization (Alter *et al.*, 2000; Hibbs *et al.*, 2005)) or clustering of the observations (Yeung and Ruzzo, 2001), with population genetics an important use case (Patterson *et al.*, 2006). PCA has also been used as the basis for feature selection (Lu *et al.*, 2011), gene clustering (Hastie *et al.*, 2000) and bi-clustering (Kluger *et al.*, 2003). More recent applications include dimensionality reduction prior to gene set testing (Tomfohr *et al.*, 2005; Kong *et al.*, 2006; Ma and Kosorok, 2009; Bruckskotten *et al.*, 2010; Chen, 2011) and high-dimensional regression (Hastie *et al.*, 2009).

Although PCA is a popular and effective tool for reducing the dimensionality of genomic data, application of the method remains limited by the challenge of biological interpretation (Zou *et al.*, 2006; Ma and Dai, 2011). Because PCs are linear combinations of all original variables, which can number from the thousands to the millions for genomic data sets, they typically lack any clear biological meaning. While PCA may improve the performance of many statistical methods, e.g., better predictive accuracy in a regression context, the underlying model is often a black box.

Approaches for generating more interpretable PCs have evolved from component thresholding (Jolliffe, 2002), simple components (i.e., PC loading vectors constrained to values from  $\{-1, 0, 1\}$ ) (Vines, 2000) and rotation techniques (e.g., varimax) (Jolliffe, 1995) to sparse PCA methods, which compute approximate PCs using cardinality (Moghaddam *et al.*, 2006; d’Aspremont *et al.*, 2007; Sriperumbudur *et al.*, 2011) or LASSO-based (Jolliffe *et al.*, 2003; Zou *et al.*, 2006; Shen and Huang, 2008; Witten *et al.*, 2009) constraints on the component loadings. By generating approximate PCs with few non-zero loadings, all of these techniques improve interpretability by associating only a small number of variables with each PC. While such sparse PCA methods can be very effective when the true population PCs are associated with only a few variables, they will fail to accurately estimate the spectral structure of the data when the population PCs are defined by the coordinated action of large groups of variables with small marginal effects. For genomic data, the pathway-based patterns that dominate the robust structure of genetic associations with clinical phenotypes (Allison *et al.*, 2006), and are the motivation for traditional gene set testing methods (Huang *et al.*, 2009; Khatri *et al.*, 2012), can be expected to also characterize the PCs of those data sets. The PCs of genomic data are therefore more likely to be quantitatively described, in a repeatable fashion, by collections of functionally related genes, e.g., gene sets from the Gene Ontology (GO) (Ashburner *et al.*, 2000) or pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), than by individual genes.

To support interpretation of PCs in terms of *a priori* variable groups, rather than just individual variables, sparse PCA methods have recently been extended to include structured sparse penalties (Jenatton *et al.*, 2010; Grbovic *et al.*, 2012), such as the group lasso (Yuan and Lin, 2006; Friedman *et al.*, 2010) and overlapping group lasso (Liu and Ye, 2010; Obozinski *et al.*, 2011). Although structured sparse PCA techniques generate sparse PC loading vectors that reflect group structure, these methods cannot be easily used to compute the statistical association between variable groups and each PC in such a way that the variable groups can be ranked according to deviation from a specific null hypothesis, as is done in traditional gene set testing. Matrix correlation methods, such as Yanai’s GCD (Yanai, 1980; Jolliffe, 2002; Ramsay *et al.*, 1984), have also been used to quantify the association between groups of variables and one or more PCs. However, because such matrix correlation methods compute the association of each variable group independent of the variables that do not belong to the group, they can only be used for self-contained

gene set tests (Goeman and Buehlmann, 2007) ( $Q_2$  in the terminology of Tian et al. (Tian *et al.*, 2005)) in a manner similar to Goeman and Buhmann’s *globaltest* (Goeman *et al.*, 2004) and not for competitive gene set testing ( $Q_1$  in the terminology of Tian et al.).

To date, competitive gene set testing relative to PCs has been limited to methods, such as Fisher’s Exact Test, that are based on a  $2 \times 2$  contingency table representing the association between gene set membership and a discretization of the ranked list of PC loading values (Roden *et al.*, 2006). Such contingency table tests have two key flaws: they rely on an arbitrary threshold of the gene-level test statistic, which reduces statistical power and, more importantly, they are based on the incorrect assumption of independence among the gene-level test statistics, causing them to generate high type I error rates (Goeman and Buehlmann, 2007; Barry *et al.*, 2008; Wu and Smyth, 2012). These same flaws apply equally in the context of gene set testing relative to PCs. Because of the anti-conservative nature of contingency table-based tests, and other approaches that assume independence among gene-level test statistics under the null, the use of these methods for standard gene set testing has been strongly discouraged in favor of techniques that preserve inter-gene correlation, usually via permutation of the sample labels (Goeman and Buehlmann, 2007). Competitive gene set testing methods that correctly account for correlation among gene-level test statistics, either through sample permutation, parametric approximation of the sample permutation distribution or correlation adjustment of parametric test statistics, include SAFE (Barry *et al.*, 2005, 2008; Zhou *et al.*, 2013), GSEA (Subramanian *et al.*, 2005), GSA (Efron and Tibshirani, 2007) and CAMERA (Wu and Smyth, 2012).

Although biologically meaningful and repeatable interpretation of the PCs of genomic data requires approaches based on functional gene sets, researchers do not currently have access to methods that competitively test the association between gene sets and PCs with correct handling of inter-gene correlation to control type I errors. To address this gap, we have developed principal component gene set enrichment (PCGSE), an approach for interpreting the PCs of genomic data via two-stage competitive gene set testing in which the correlation between each gene and each PC is used as a gene-level statistic with flexible choice of both the gene set test statistic and the method used to compute the null distribution of the gene set statistic. Although described in the context of functional gene sets and genomic data, the PCGSE method can be used to compute the statistical association between any collection of variable groups and the PCs of an empirical dataset. To enable the easy application of the PCGSE method by other researchers, we implemented the *PCGSE* R package, which can be downloaded from the CRAN repository. Using simulated data with simulated gene sets and real gene expression data with curated gene sets, we demonstrate that biologically meaningful and computationally efficient results can be obtained from a simple parametric version of the PCGSE technique, based on the CAMERA method (Wu and Smyth, 2012), that performs a correlation-adjusted two-sample t-test between the gene-level test statistics for gene set members and genes not in the set.

## 2 Methods

### 2.1 PCGSE inputs

The PCGSE method takes the following data structures as input:

1. *Matrix of genomic data:*  $n \times p$  matrix  $\mathbf{X}$  quantifying  $p$  genomic variables under  $n$  experimental conditions, e.g., mRNA expression levels measuring using microarray technology. This data will be modeled as a sample of  $n$  independent observations from a  $p$ -dimensional random vector  $\mathbf{x}$ . Although PCGSE does not have specific distributional requirements, sources of genomic

data, especially gene expression data, are typically well represented by a multivariate normal distribution  $\sim \mathcal{N}(\mu_{p \times 1}, \Sigma_{p \times p})$ , especially after appropriate transformations. Often,  $p \gg n$ .

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \quad (1)$$

where  $x_{i,j}$  represents the abundance of genomic variable  $j$  under condition  $i$ . It is assumed that any desired data transformations (e.g., log transformation of mRNA expression ratios, etc.) have been performed and that missing values have been imputed or removed for a complete case analysis.

2. *Matrix of functional annotations:*  $f \times p$  binary annotation matrix  $\mathbf{A}$  whose rows represent  $f$  different biological functions, e.g., GO terms or KEGG pathways, and whose cells  $a_{i,j}$  hold indicator variables whose value depends on whether an annotation exists between the function  $i$  and genomic variable  $j$ .

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{f,1} & \cdots & a_{f,p} \end{bmatrix}, a_{i,j} = \mathbb{1}[\text{variable } j \text{ has function } i] \quad (2)$$

3. *Algorithm parameters:* The PCGSE method requires the specification of parameters that determine the gene-level statistic used to quantify association between genomic variables and PCs, whether transformations are applied to the gene-level statistics, the type of statistic calculated for each gene set and the method used to assess the significance of the gene set statistic given a competitive null hypothesis.

## 2.2 PCGSE algorithm

Enrichment of the gene sets defined by  $\mathbf{A}$  relative to one of the PCs of  $\mathbf{X}$  is performed using the following sequence of steps:

1. Perform PCA on a standardized version of  $\mathbf{X}$ .
2. Compute gene-level statistics,  $z_j, j = 1, \dots, p$ , for all  $p$  genomic variables that quantify the association between the genomic variable and the PC.
3. (Optional) Transform the gene-level statistics.
4. Compute gene set statistics,  $S_k, k = 1, \dots, f$ , for all  $f$  gene sets defined by  $\mathbf{A}$  using the gene-level statistics,  $z_j$ .
5. Determine the statistical significance of the gene set statistics according to a competitive null hypothesis.

Each of these steps is explained in more detail in the sections 2.3 thru 2.7 below. Note that steps 2 thru 5 have close parallels to modules in Ackermann and Strimmer's general modular framework for gene set enrichment analysis (Ackermann and Strimmer, 2009).

### 2.3 PCA for PCGSE

Because PCs are not invariant under scaling of the data (Jolliffe, 2002), PCA is performed on a mean centered and standardized version of  $\mathbf{X}$ :

$$\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{x}_{1,1} & \cdots & \tilde{x}_{1,p} \\ \vdots & \ddots & \vdots \\ \tilde{x}_{n,1} & \cdots & \tilde{x}_{n,p} \end{bmatrix}, \hat{x}_{i,j} = \frac{x_{i,j} - \bar{x}_j}{s_{x_j}} \quad (3)$$

where  $\bar{x}_j$  is the mean value of the  $j$ th genomic variable computed over the  $n$  samples and  $s_{x_j}$  is the sample standard deviation of the  $j$ th genomic variable. The PC loading vectors and variances of  $\tilde{\mathbf{X}}$  are thus the eigenvectors and eigenvalues of the sample correlation matrix,  $\mathbf{S} = 1/(n-1)\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ , rather than the sample covariance matrix. For computational efficiency, the PCA solution is realized via the singular value decomposition (SVD) of  $\tilde{\mathbf{X}}$ :

$$\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (4)$$

where the columns of  $\mathbf{V}$  represent the PC loading vectors, the entries in the diagonal matrix  $\mathbf{\Sigma}$  are proportional to the square roots of the PC variances and the columns of  $\mathbf{U}\mathbf{\Sigma}$  are the PCs.

### 2.4 Gene-level statistics

The PCGSE method supports the following gene-level statistics for quantifying the association between genomic variable  $j$  and the target PC. These statistics are represented using the notation  $z_j, j = 1, \dots, p$ .

- *PC loading.* For genomic variable  $j$  and target PC  $m$ , the gene-level statistic is element  $v_{j,m}$  of matrix  $\mathbf{V}$  from the SVD of  $\tilde{\mathbf{X}}$  as defined in (4).
- *Pearson correlation coefficient.* Where the correlation is computed between each genomic variable and the target PC.
- *Fisher-transformed Pearson correlation coefficient.* This creates a statistic whose distribution is approximately  $\mathcal{N}(0, 1)$ .

Because the Pearson correlation coefficients between genomic variables and PCs of the sample correlation matrix are proportional to the PC loadings (see (5) below), all of these gene-level statistics provide a measure of the correlation between genomic variables and PCs.

$$\begin{aligned} \text{cor}(\tilde{\mathbf{X}}, \mathbf{U}\mathbf{\Sigma}) &= \text{cov}(\tilde{\mathbf{X}}, \mathbf{U}\mathbf{\Sigma}\sqrt{1-n}\mathbf{\Sigma}^{-1}) \\ &= \frac{1}{n-1}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T\mathbf{U}\sqrt{1-n} \\ &= \frac{\sqrt{n-1}}{n-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U} = \frac{1}{\sqrt{n-1}}\mathbf{V}\mathbf{\Sigma} \end{aligned} \quad (5)$$

where  $\mathbf{U}$ ,  $\mathbf{\Sigma}$  and  $\mathbf{V}$  are from the SVD of  $\tilde{\mathbf{X}}$  as specified in (4).

The choice between the different gene-level statistics will be guided by the gene set statistic and significance testing method employed for PCGSE as well as computational constraints. For example, the added computational expense to generate z-statistics from correlation coefficients is motivated by parametric tests of the mean difference statistic, whereas, for rank sum tests, the PC loadings are sufficient.

## 2.5 Transformation of gene-level statistics

An absolute value transformation can optionally be applied to the gene-level statistics, i.e.,  $\tilde{z}_j = |z_j|$ . Such a transformation gives the PCGSE method increased power to detect scale alternatives, i.e. gene sets that contain both significantly enriched and significantly repressed genomic variables, whereas the use of untransformed gene-level statistics provides better power against shift in location alternatives, i.e., gene sets containing genomic variables with a common direction of association (Efron and Tibshirani, 2007).

## 2.6 Gene set statistics

The PCGSE method supports two competitive gene set statistics for quantifying the association between gene set  $k$  and a target PC. These statistics are represented using the notation  $S_k, k = 1, \dots, f$ .

### 2.6.1 Mean difference statistic

This statistic is computed as the standardized difference between the mean of the  $z_j$  for genomic variables in the gene set and genomic variables not in the set and corresponds to  $U_D$  in the notation of Barry *et al.* (2008). Benefits of the mean difference statistic include its parametric null distribution and excellent power, relative to other gene set test statistics, for shift in location alternatives when using untransformed  $z_j$  (Efron and Tibshirani, 2007). For gene set  $k$ , this statistic is defined as:

$$S_k = \frac{\bar{z}_k - \bar{z}_{k^c}}{\sigma_p \sqrt{\frac{1}{m_k} - \frac{1}{p-m_k}}} \quad (6)$$

$$m_k = \sum_{j=1}^p a_{k,j}, \bar{z}_k = \frac{\sum_{j=1}^p a_{k,j} z_j}{m_k}, \bar{z}_{k^c} = \frac{\sum_{j=1}^p !a_{k,j} z_j}{p - m_k}$$

where  $m_k$  is the number of genes in set  $k$ ,  $\bar{z}_k$  is the mean of the  $z_j$  for members of gene set  $k$ ,  $\bar{z}_{k^c}$  is the mean of the  $z_j$  for genes not in set  $k$  and  $\sigma_p$  is the pooled standard deviation of the  $z_j$ .

### 2.6.2 Rank sum statistic

This statistic is computed as the standardized Wilcoxon rank sum statistic given the ranks of the  $z_j$  for genomic variables in the set and genomic variables not in the set and corresponds to  $U_W$  in the notation of Barry *et al.* (2008). Benefits of the rank sum statistic include lack of distributional assumptions and robustness to outliers. For gene set  $k$ , the Wilcoxon rank sum statistic is defined as the sum of the ranks of the gene-level statistic for all genomic variables belonging to gene set  $k$  minus the minimum possible value for this sum of ranks:

$$W_k = \sum_{j=1}^p a_{k,j} \text{Rank}(z_j) - \frac{m_k(m_k + 1)}{2} \quad (7)$$

where  $m_k = \sum_{j=1}^p a_{k,j}$ , the size of gene set  $k$ . A version of this statistic that has an asymptotic  $\mathcal{N}(0, 1)$  distribution under the null can be generated as:

$$S_k = \frac{W_k - \mu_{W_k}}{\sigma_{W_k}^2} \quad (8)$$

where  $\mu_{W_k} = (m_k(p - m_k))/2$  and  $\sigma_{W_k}^2 = (m_k(p - m_k)(m_k + 1))/12$ .

Although it is possible to use any of the supported gene-level statistics with either of the gene set statistics, as discussed above, the mean difference statistic should be used with Fisher-transformed Pearson correlation coefficients and the rank sum statistic can be effectively used with just the PC loading elements.

## 2.7 Gene set statistical significance

To compute the statistical significance of the association between gene set  $k$  and a target PC, the distribution of the gene set statistic  $S_k$  must be calculated under the appropriate null hypothesis. The PCGSE approach supports three different methods (parametric, correlation-adjusted parametric and permutation) for computing the competitive null distributions of the mean difference and rank sum gene set statistics defined in (6) and (8).

### 2.7.1 Parametric tests

Under the competitive  $H_0$  that the  $z_j$  are independent and identically distributed, it is possible to determine the statistical significance of the association between each gene set and the target PC using a two-sided t-test for the mean difference statistic or a two-sided z-test for the rank sum statistic. Both of these parametric tests fall into the class 1 test category as outlined in Barry *et al.* (2008) and are similar to the  $Q_1$  test defined by Tian *et al.* (2005).

Under this  $H_0$ , the mean difference statistic defined in (6) has a t-distribution with  $p - 2$  df and a two-sided t-test can therefore be used to determine statistical significance. For the rank sum statistic defined in (8), the asymptotic standard normal distribution under this  $H_0$  can be used as the basis for a two-sided z-test.

While it is often safe to assume a normal distribution for the  $z_j$ , especially after transformation, the  $z_j$  will not be independent. Indeed, because the  $z_j$  used with PCGSE are proportional to the PC loadings, they have an asymptotic multivariate normal distribution (Anderson, 1963), assuming multivariate normality for the underlying genomic data, with significant correlation present between the loadings associated with the genes that have high pair-wise correlations (Jolliffe, 2002). Because both the t-test for the mean difference statistic and the z-test for the rank sum statistic ignore this correlation between gene-level statistics, they will generate inflated type I error rates. These tests are therefore only supported by the PCGSE method for the purpose of comparative evaluation.

### 2.7.2 Correlation-adjusted parametric tests

A computationally efficient approach for addressing correlation among the  $z_j$  involves the use of correlation-adjusted parametric tests. Correlation-adjusted versions of the t-statistic associated with the mean difference statistic defined in (6) and of the z-statistic associated with the rank sum statistic defined in (8) were first discussed in the context of gene set testing by Barry *et al.* (2008). Simplified versions of these correlation-adjusted statistics were later developed into the CAMERA method by Wu and Smyth (2012). Specifically, the approach taken by CAMERA assumes that correlation among the  $z_j$  can be approximated by the correlation among the genomic variables (this is supported by results in Barry *et al.* (2008)), ignores all inter-gene correlation except the correlation among the members of the tested gene set and estimates a single average pair-wise correlation for gene set members using residuals from a linear regression.

The PCGSE method makes similar simplifying assumptions as those made by CAMERA, i.e., correlation between the  $z_j$  can be approximated by correlation among the genomic variables, only gene set members have non-zero inter-gene correlation and all pair-wise correlations between

gene set members are the same. An important difference between PCGSE and CAMERA is that PCGSE estimates the average inter-gene correlation directly from the sample correlation matrix. The correlation-adjusted mean difference statistic used by PCGSE is:

$$S_k^{adj} = \frac{\bar{z}_k - \bar{z}_{k^c}}{\sigma_p \sqrt{\frac{VIF}{m_k} - \frac{1}{p-m_k}}} \quad (9)$$

where  $VIF$  (variance inflation factor) =  $1 + (m_k - 1)\bar{\rho}_k$  and  $\bar{\rho}_k$  is the average unbiased sample correlation between members of gene set  $k$ . Following Wu and Smyth (2012), this correlation-adjusted statistic has a t-distribution with  $n - 2$  df under  $H_0$ . Likewise the correlation-adjusted rank sum statistic is computed as:

$$S_k^{adj} = \frac{W_k - \mu_{W_k}}{\sigma_{VIF, W_k}^2} \quad (10)$$

where  $\sigma_{VIF, W_k}^2 = (m_k(p - m_k)) / (2\pi)(\sin^{-1}(1) + (p - m_k - 1)\sin^{-1}(.5) + (m_k - 1)(p - m_k - 1)\sin^{-1}(\bar{\rho}_k/2) + (m_k - 1)\sin^{-1}((\bar{\rho}_k + 1)/2))$ , as derived in Wu and Smyth (2012) based on the formula in Barry *et al.* (2008).

### 2.7.3 Permutation test

The most common approach in the gene set testing literature for addressing correlation between gene-level statistics has been sample permutation. This approach, which corresponds to the class 2 test in Barry *et al.* (2008), generates the null distribution of the gene set statistic via permutation of the outcome variable. For each permutation of the outcome variable, all gene-level statistics are recomputed to generate permutation statistics  $z_j^*$  and then permutation gene set statistics  $S_k^*$  are calculated using the  $z_j^*$ . The statistical significance for a given gene set  $k$  is based on the proportion of all permutation  $S_k^*$  more extreme than the observed  $S_k$ . In standard gene set testing, permutation is applied to a clinical outcome variable, e.g., a case/control label. For PCGSE, permutation is applied to the elements of the target PC, i.e., the elements of one of the columns of  $\mathbf{U}\Sigma$ . Because permutation is applied to the PC elements, this test can only be used with Pearson correlation coefficients or Fisher-transformed Pearson correlation coefficients as gene-level statistics.

A key assumption of the permutation null distribution is that the permuted values are i.i.d. Assuming the original  $n$  observations of the  $p$ -dimensional random vector  $\mathbf{x}$  are i.i.d, the elements of each PC will also be i.i.d., since each PC is a linear function of the original  $\mathbf{x}$ . Permutation of the PC elements therefore generates a valid permutation distribution for the mean difference and rank sum gene set statistics.

Because permutation tests handle correlation among the  $z_j$  without attempting to estimate this correlation or make simplifying assumptions about the correlation structure, they are likely the most accurate of the statistical tests supported by PCGSE and are therefore used to evaluate the performance of the parametric and correlation-adjusted parametric tests. The exact permutation test was also used as a "gold-standard" in Zhou *et al.* (2013). Although they provide superior handling of inter-gene correlation, permutation tests do suffer from two important disadvantages relative to parametric tests: computational complexity and lower power to detect gene sets whose members all have a small common association with the outcome. Because of these disadvantages, correlation-adjusted parametric tests are preferred for most PCGSE applications.

Another alternative to sample permutation testing that addresses the key challenge of computational complexity is the parametric approximation of the sample permutation distribution of gene-level score statistics developed by Zhou *et al.* (2013). Although the Zhou *et al.*'s beta



distribution-based parametric approximations may be a useful option for the PCGSE method, it is not currently supported due to the lack of a parametric approximation for a directional, competitive gene set test statistic that is equivalent to the standardized mean difference statistic using untransformed directional gene-level test statistics. In Zhou *et al.* (2013), parametric approximations are only detailed for two self-contained gene set test statistics (sum of the score statistics and sum of the squares of the score statistics) and one non-directional competitive test statistic (a weighted sum of the squares of local score statistics).

## 2.8 PCGSE output

For each of the  $f$  gene sets defined in  $\mathbf{A}$  and each tested PC of  $\tilde{\mathbf{X}}$ , the PCGSE method outputs the observed value of the gene set test statistic,  $S_k$ , and a p-value representing the probability of encountering a gene set statistic as or more extreme than the observed  $S_k$  under the appropriate competitive null hypothesis.

## 2.9 PCGSE evaluation

### 2.9.1 Evaluation using simulated gene sets and simulated data.

As a simple example, the PCGSE method was used to compute the statistical association between 20 disjoint gene sets, each of size 10, against the PCs of 100 simulated gene expression datasets each comprised by 50 independent observations of a 200-dimensional random vector simulated according to a multivariate normal distribution  $\sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . The population covariance matrix,  $\boldsymbol{\Sigma}$ , was generated as:  $\boldsymbol{\Sigma} = \lambda_1 \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^T + \lambda_2 \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_2^T + \lambda_d \mathbf{I}$ , where  $\lambda_1 = 2$ ,  $\lambda_2 = 1$ ,  $\lambda_d = 0.1$ ,  $\boldsymbol{\alpha}_1$  is a 200-dimensional vector with all elements equal to 0 except for the first 10 which were set to  $\sqrt{1}$ ,  $\boldsymbol{\alpha}_2$  is a 200-dimensional vector with all elements equal to 0 except for the second 10 which were set to  $\sqrt{1}$ . Figure 1 shows the variance and loadings for both population and sample PCs simulated according to this model.

For this simulated example, the PCGSE method was executed using the Fisher-transformed Pearson correlation coefficient between each variable and each PC as the gene-level test statistic with the standardized mean difference, as defined in (6), as the gene set test statistic. The statistical significance of the association between each of the 20 simulated gene sets and each PC was computed using all supported tests described in Section 2.7: parametric, correlation-adjusted parametric and permutation. For the standardized mean difference gene set statistic, these tests were realized by a two-sided t-test, a correlation-adjusted two-sided t-test and a two-sided permutation test based on permutation of the PC elements, respective.

Because the true association was known between simulated gene sets and the PCs of the simulated data, it was possible to compute contingency table statistics. In this case, the type I error rates for the different statistical testing methods were computed for gene set 2 relative to PC 1 and for gene set 1 relative to PC 2, both cases with no true association.

### 2.9.2 Evaluation using Spellman *et al.* $\alpha$ factor-synchronized yeast gene expression data and yeast cell cycle gene sets.

The PCGSE method was used to compute the statistical association of the yeast cell cycle gene sets defined by Spellman (Spellman *et al.*, 1998) relative to the first three PCs of a specially processed version of the  $\alpha$  factor-synchronized yeast gene expression data collected by Spellman *et al.* (1998) and re-examined by Alter *et al.* (2000).

Both the  $\alpha$  factor-synchronized data and yeast cell cycle gene sets were downloaded from the supplementary material website for Alter *et al.* (2000). To support comparison against the results reported in Alter *et al.* (2000), PCA was performed on a version of the gene expression data that was specially processed according to the steps outlined in Alter *et al.* (2000) so that the first three PCs were identical to the first three so-called eigengenes. Figure 3 is a reproduction of Figure 5 from Alter *et al.* (2000) with the value of the first three PCs of the specially processed gene expression data (i.e., the eigengenes) shown relative to the 22  $\alpha$  factor arrays.

The PCGSE method was executed on the Spellman *et al.* data and gene sets using the Fisher-transformed Pearson correlation coefficient between each gene and each PC as the gene-level test statistic and the standardized mean difference, as defined in (6), as the gene set statistic. Similar to the simulation example outlined in Section 2.9.1, the statistical significance of the gene set statistic was computed using all supported tests described in Section 2.7.

### 2.9.3 Evaluation using MSigDB C2 v4.0 gene sets and Armstrong *et al.* leukemia gene expression data.

The PCGSE method was also used to compute the statistical association between the MSigDB C2 v4.0 gene sets and the first 3 PCs of the leukemia gene expression data (Armstrong *et al.*, 2002) used in the 2005 GSEA paper (Subramanian *et al.*, 2005).

The MSigDB C2 v4.0 cancer modules and collapsed leukemia gene expression data were both downloaded from the MSigDB repository. With a minimum gene set size of 15 and maximum gene set size of 200, 3,076 gene sets out of the original 4,722 were used in the analysis. Similar to the simulation example outlined in Section 2.9.1 and the yeast cell cycle example outlined in Section 2.9.2, the PCGSE method was executed using the Fisher-transformed Pearson correlation coefficient between each genomic variable and each PC as the gene-level test statistic and the standardized mean difference, as defined in (6), as the gene set test statistic. The statistical significance of the association between each of the MSigDB C2 gene sets and each of the first 3 PCs of the standardized leukemia gene expression data was computed using all supported tests described in Section 2.7.

The enrichment of the MSigDB C2 gene sets was also computed relative to the acute myeloid leukemia (AML) versus acute lymphoblastic leukemia (ALL) phenotype using the GSA method (Efron and Tibshirani, 2007) with the restandardized mean statistic and 10,000 permutations. For each of the first three PCs and each of the PCGSE methods for computing statistical significance of the standardized mean difference gene set statistic, the Spearman correlation coefficient was computed between PC gene set enrichment p-values and phenotype enrichment p-values. For PC 2, for which the PC and phenotype gene set enrichment p-values were highly correlated, contingency table statistics were computed measuring how well PCGSE was able to identify MSigDB C2 gene sets significantly associated with the AML/ALL phenotype.

## 3 Results and Discussion

### 3.1 Enrichment of simulated gene sets relative to PCs of simulated data

According to the population covariance matrix,  $\Sigma$ , used to simulate the 100 datasets, only the first gene set should be significantly enriched on the first PC and only the second gene set should be significantly enriched on the second PC. This relationship can be seen easily in the loading values for population PCs 1 and 2 as shown in Figure 1 plots (c) and (e). The significant loading of gene set 2 on PC 2, however, will result in a high pair-wise correlation between the PC loadings for gene set 2 members on PC 1. The fact that high loadings on one PC result in correlation among the PC

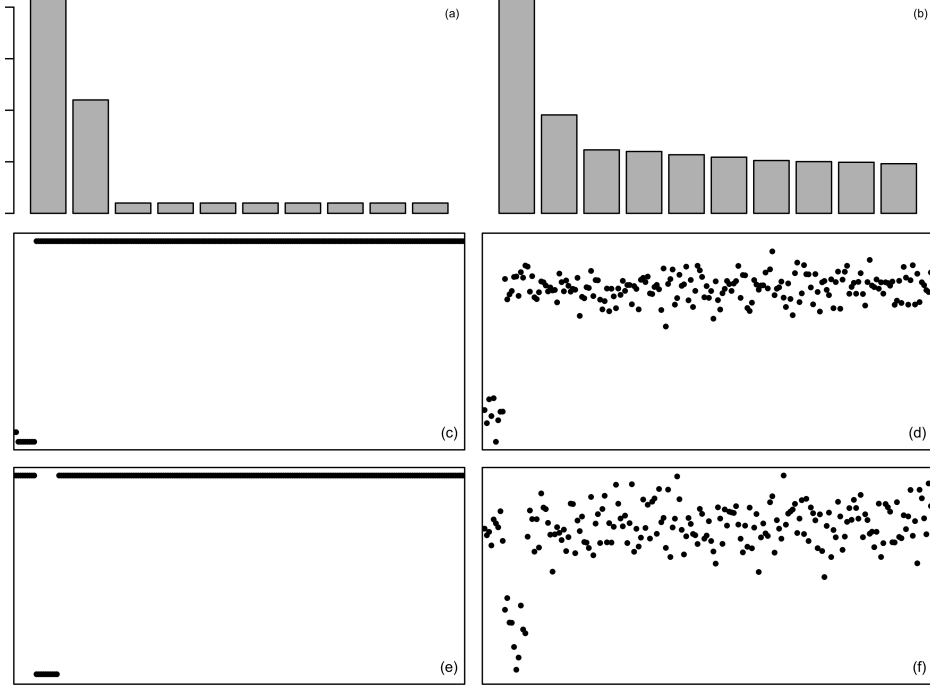


Figure 1: Simulation model. Variances and loadings for the principal components a 200-dimensional population covariance matrix,  $\Sigma$ , and the sample covariance matrix estimated from  $n=50$  independent observations of the random vector  $\mathbf{x} \sim MVN(\mathbf{0}, \Sigma)$  where  $\Sigma$  is generated according to the model outlined in Section 2.9.1. Variances for the first ten population PCs are shown in plot (a) and loadings for the first two population PCs are shown in plots (c) and (e). Plots (b), (d) and (f) show the corresponding variances and loadings for the sample PCs of a single simulated dataset.

loadings on other PCs follows from the formula for the asymptotic distribution of the PC loadings for MVN data (Anderson, 1963):

$$\mathbf{v}_j \sim \mathcal{N}(\boldsymbol{\alpha}_j, \mathbf{T}_j), j = 1, \dots, p \quad (11)$$

$$\mathbf{T}_j = \frac{\lambda_j}{n-1} \sum_{k=1, k \neq j}^p \frac{\lambda_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T}{(\lambda_k - \lambda_j)} \quad (12)$$

where

- $p$  is fixed and  $n \rightarrow \infty$ .
- $\lambda_j$  an eigenvalue of the population covariance matrix,  $\lambda_1 > \lambda_2 > \dots > \lambda_p$ .
- $\boldsymbol{\alpha}_j$  is an eigenvector of the population covariance matrix.

The gene-level test statistics computed for gene set 2 on PC 1 and for gene set 1 on PC 2 will therefore have a non-zero average pair-wise correlation. The impact of this correlation between the gene-level test statistics can be seen in the PCGSE results shown in Figure 2. The unadjusted t-test uses an incorrectly small variance for the mean difference statistic and, as expected, generates the high type I error rate of 0.42 given a nominal  $\alpha$  of 0.05 for gene set 2 relative to PC 1 and

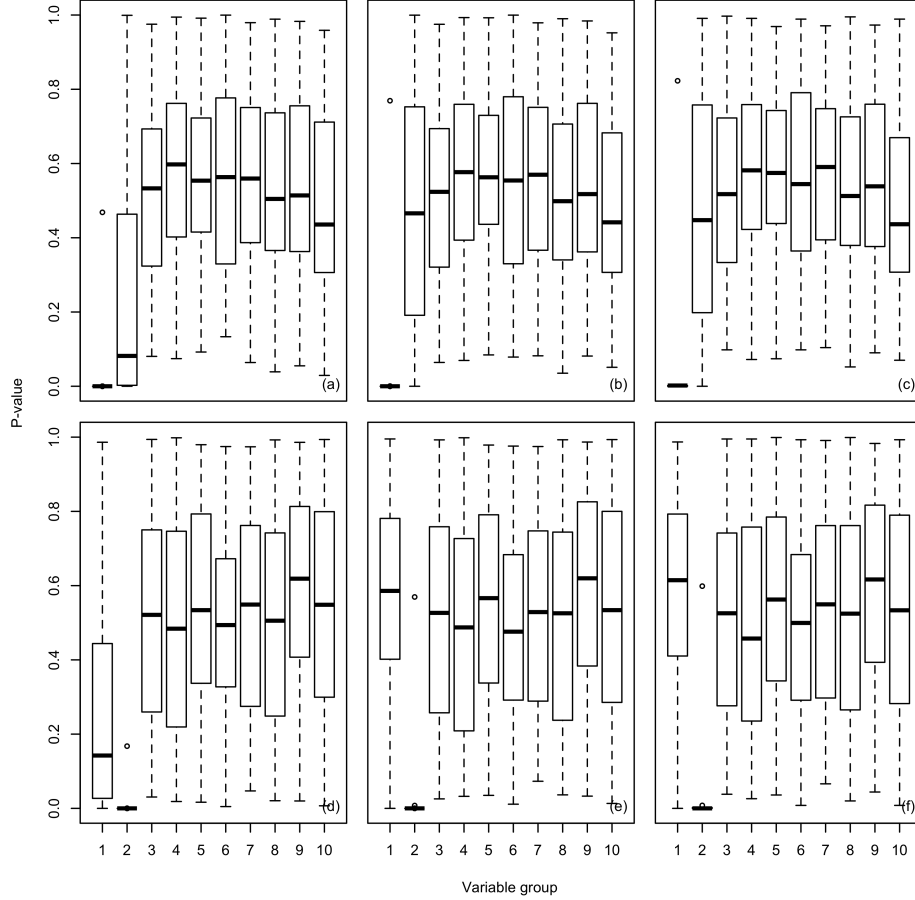


Figure 2: Simulation results. Distribution of PCGSE-computed enrichment p-values for the first 10 of 20 simulated gene sets relative to the first 2 PCs of 100 datasets simulated according to the model described in Section 2.9.1 and illustrated in Figure 1. The PCGSE method was executed using parameter settings outlined in Section 2.9.1. For all displayed results, PCGSE was executed using the Fisher-transformed Pearson correlation coefficient between each genomic variable and each PC as the gene-level test statistic and the standardized mean difference as the gene set test statistic. Plots (a), (b) and (c) display the distribution of enrichment p-values for the first 10 gene sets relative to the first PC of all simulated data sets. In plots (d), (e) and (f), enrichment p-values computed relative to the second PC are displayed. For plots (a) and (d), the p-values were computed using a two-sided t-test on the standardized mean difference gene set test statistic, for plots (b) and (e), the p-values were computed using a two-sided correlation-adjusted t-test and, for plots (c) and (f), the p-values were computed using a two-sided permutation test.

0.3 for gene set 1 relative to PC 2. The correlation-adjusted two-sided t-test and the two-sided permutation test are much more successful at controlling the type I error rate. For PC 1 and gene set 2, the type I error rate was 0.13 for both correlation-adjusted t-test and the permutation test. For PC 2 and gene set 1, the type I error rate was 0.06 for both the correlation-adjusted t-test and the permutation test. For this example, all gene set testing methods were able to correctly reject the null hypothesis for almost all cases where the gene set had a true association with the PC, e.g., gene set 1 relative to PC 1 and gene set 2 relative to PC 2.

Although based on a simple two-factor MVN model, this simulation example demonstrates the importance of controlling for correlation between gene-level test statistics when computing PC gene set enrichment. Tests which assume independence among the statistics that quantify the association between genes and PCs, such as a two-sample t-test, Fisher’s exact test or a gene permutation test, will underestimate the variance of the gene set test statistic and therefore reject too many null hypotheses. This example also shows that the correlation-adjusted t-test can achieve enrichment sensitivity and specificity comparable to a sample permutation test with a significantly lower computational burden.

PCGSE was computed for this simulation example using the standardized rank sum as the gene set statistic. The results for the standardized rank sum statistic using parametric and permutation tests are similar to those for the standardized mean difference statistic. Although the correlation-adjusted z-test based on the standardized rank sum statistic has an improved type I error rate, it has an inflated type II error rate, i.e., it rejects too few null hypotheses when the association is true. The inflated type II error rate for the correlation-adjusted rank sum z-test is likely due to an overestimated VIF under the alternative hypothesis, as computed by the equation from Barry *et al.* (2008). Therefore, in cases where a rank sum gene set statistic is motivated, PCGSE should be performed using a permutation test and not using the more efficient correlation-adjusted z-test.

### 3.2 Enrichment of Spellman *et al.* yeast cell cycle gene sets relative to the PCs of Spellman *et al.* $\alpha$ -synchronized yeast gene expression data

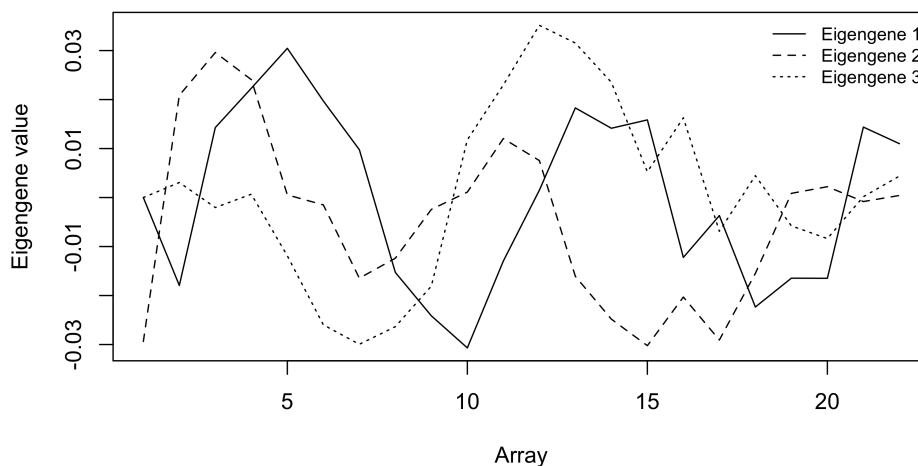


Figure 3: Reproduction of Figure 5 from Alter *et al.* (2000) that displays the value of first three PCs of the specially processed gene expression matrix, i.e., the “eigengenes”, over the 22  $\alpha$  factor-synchronized arrays. According to Alter *et al.* (2000), the following approximate mapping held between arrays and cell cycle phases: 1 (NA), 2 (M/G1), 3,4 (G1), 5,6 (S), 7,8 (S/G2), 9 (G2/M), 10,11 (M/G1), 12,13 (G1), 14,15 (S), 16 (S/G2), 17,18 (G2/M), 19,20 (S/G1), 21,22 (G1)

The Spellman *et al.* (1998)  $\alpha$  factor-synchronized gene expression data was selected for PCGSE analysis because it is well known, has been widely reanalyzed, is easily accessible and has a spectra with a published biological interpretation. In particular, the reanalysis by Alter *et al.* (2000) was one of the first to illustrate that the spectra of gene expression data can represent important biological features, in this case phases of the yeast cell cycle. In Alter *et al.* (2000), the authors

	T-test		Cor-adj t-test		Perm	
	PC 1	PC 2	PC 1	PC 2	PC 1	PC 2
$M/G_1$	0.68	<b>1.2e-12</b>	0.94	0.18	0.94	0.22
$G_1$	<b>3.5e-130</b>	<b>1e-35</b>	<b>0.023</b>	0.23	<b>0.024</b>	0.35
$S$	<b>1e-10</b>	<b>0.0074</b>	0.2	0.59	0.2	0.62
$S/G_2$	0.27	<b>4.6e-06</b>	0.86	0.45	0.87	0.47
$G_2/M$	<b>8.3e-38</b>	0.068	0.07	0.79	<b>0.048</b>	0.81

Table 1: PCGSE computed enrichment p-values for the Spellman *et al.* (1998) yeast cell cycle gene sets relative to the first two PCs of the Spellman *et al.* (1998)  $\alpha$  factor-synchronized gene expression data processed using the steps outlined in Alter *et al.* (2000). PCGSE was executed using Fisher transformed Pearson correlation coefficients between genes and PCs as gene-level test statistics. Significance of the standardized mean difference gene set statistic was computed using either a two-sided t-test, a correlation-adjusted two-sided t-test or a two-sided permutation test. Unadjusted p-values less than 0.05 are displayed in bold.

provided a qualitative interpretation of the first two eigengenes in terms of the yeast cell cycle by examining the correlation between the eigengenes and genes known to be active during different cell cycle phases, as defined by Spellman *et al.*'s yeast cell cycle gene sets. Alter *et al.* concluded that the first eigengene was correlated with genes that peak late in cell cycle phase  $G_1$  and early in phase  $S$  and was anticorrelated with genes that peak late in cell cycle phase  $G_2/M$  and early in phase  $M/G_1$ . Alter *et al.* also concluded that the second eigengene was correlated with genes that peak late in cell cycle phase  $M/G_1$  and early in phase  $G_1$  and was anticorrelated with genes that peak late in phase  $S$  and early in phase  $S/G_2$ .

Table 3.2 contains p-values representing the statistical significance of the association between each of the Spellman *et al.* yeast cell cycle gene sets and the first two PCs of a specially processed version of the Spellman *et al.* gene expression data. As described in Section 2.9.2, this special processing ensured that the PCs were identical to the eigengenes analyzed in Alter *et al.* (2000). When a unadjusted two-sided t-test was used to determine the statistical significance of the standardized mean difference gene set statistic, the gene sets corresponding to cell cycles  $G_1$ ,  $S$  and  $G_2/M$  were all highly significantly associated with PC 1 and the gene sets corresponding to  $M/G_1$ ,  $G_1$ ,  $S$  and  $S/G_2$  were all significantly associated with PC 2. However, when either a correlation-adjusted two-sided t-test or two-sided permutation was used to determine the statistical significance of the standardized mean difference set statistic, PC 1 only had a significant association with the gene set corresponding to phase  $G_1$  (with a marginally significant association with phase  $G_2/M$ ) and none of the cell cycle gene sets were significantly associated with PC 2.

Comparing the output from PCGSE with the analysis in Alter *et al.* (2000), the results from the unadjusted two-sided t-test align closely with the qualitative conclusions of Alter *et al.* The output from the correlation-adjusted t-test and permutation test, although generally in agreement for PC 1, are in direct contract with Alter *et al.* regarding PC 2, finding no cell cycle association. The agreement between Alter *et al.* and the unadjusted t-test results is expected since the authors had based their analysis simply on a qualitative inspection of the gene-level correlations without a more formal test of a gene set test statistic that took account of the correlation between the gene-level test statistics associated with each cell cycle phase. The fact that the more accurate PCGSE methods failed to find an association between PC 2 and the cell cycle gene sets indicates that the originally published association in Alter *et al.* (2000) was a false positive due to either the high inter-gene correlation present among the members of these sets or the selective examination

by Alter et al. on a subset of the genes in each of the cell cycle gene set with a common direction of association with the eigengene. In the later case, it is likely that a gene set statistic such as the maxmean (Efron and Tibshirani, 2007) would identify significant cell cycle enrichment for the second eigengene.

This example highlights the importance of using formal statistical methods for gene set testing when attempting to interpret the PCs of genomic data in terms of gene sets. Such gene set testing methods must specifically account for the correlation between gene-level test statistics.

### 3.3 Enrichment of MSigDB C2 v4.0 gene sets relative to PCs of Armstrong et al. leukemia gene expression data

The classic Armstrong *et al.* (2002) leukemia gene expression dataset is another excellent example of a case where the genomic patterns associated with an interesting phenotype have a clear representation in the spectral structure of the data. For the Armstrong et al. data, the second PC of the gene expression data is strongly associated with the AML versus ALL status of the subjects. Use of the Armstrong et al. gene expression data and MSigDB C2 v4.0 gene sets for evaluation of PCGSE was also motivated by the extensive use of this dataset and gene set collection in the gene set enrichment literature (e.g., Subramanian *et al.* (2005)) and easy accessibility from the MSigDB repository, factors that will facilitate interpretation and replication of the reported PCGSE results by other researchers.

Figure 4 shows the association between phenotype and PC gene set enrichment p-values for the MSigDB C2 v4.0 gene sets, the AML versus ALL phenotype and the first three PCs of the Armstrong et al. leukemia gene expression data. Each of the columns in the multi-plot corresponds to results for one PC and each row corresponds to one of the three different statistical tests supported by PCGSE on the standardized mean difference gene set statistic (i.e., t-test, correlation-adjusted t-test and permutation test). The association between PC 2 and the AML versus ALL phenotype can be clearly seen in Figure 4 plots (b), (e) and (h). For all three PCGSE methods, the PC enrichment p-values for the MSigDB C2 v4.0 gene sets are highly correlated with the enrichment p-values computed for these gene sets relative to the AML versus ALL phenotype.

Similar to the PCGSE results outlined in previous sections on simulated data and yeast gene expression data, the unadjusted two-sided t-test on the standardized mean difference gene set statistic generates PC gene set enrichment p-values that are substantially lower than the enrichment p-values output by either the correlation-adjusted t-test or the permutation test. Although the true enrichment status of the MSigDB C2 v4.0 gene sets relative to the PCs of the Armstrong *et al.* (2002) gene expression data is unknown, the phenotype enrichment results can be used as a proxy for the true gene set association with PC 2 under the assumption that this PC captures the AML versus ALL signal. If gene sets with a phenotype enrichment significance at or below 0.05 are considered AML/ALL markers, the PCGSE method is able to correctly identify these gene sets via enrichment relative to PC 2 with an area under the receiver operator characteristic curve (AUC) of 0.85 for the t-test results displayed in plot (b), an AUC of 0.89 for the correlation-adjusted t-test results displayed in plot (e) and an AUC of 0.85 for the permutation test results displayed in plot (h). Considering identification of AML/ALL-associated gene sets via PC enrichment using just  $\alpha = 0.05$ , the PCGSE method has a positive predictive value of 0.26 for the t-test results displayed in plot (b), 0.91 for the correlation-adjusted t-test results displayed in plot (e) and 0.46 for the permutation test results displayed in plot (h).

PCGSE analysis of the MSigDB C2 v4.0 gene sets and Armstrong *et al.* (2002) leukemia gene expression data illustrates the biological motivation for PC gene set enrichment and demonstrates the superior performance of the computationally efficient correlation-adjusted t-test relative to

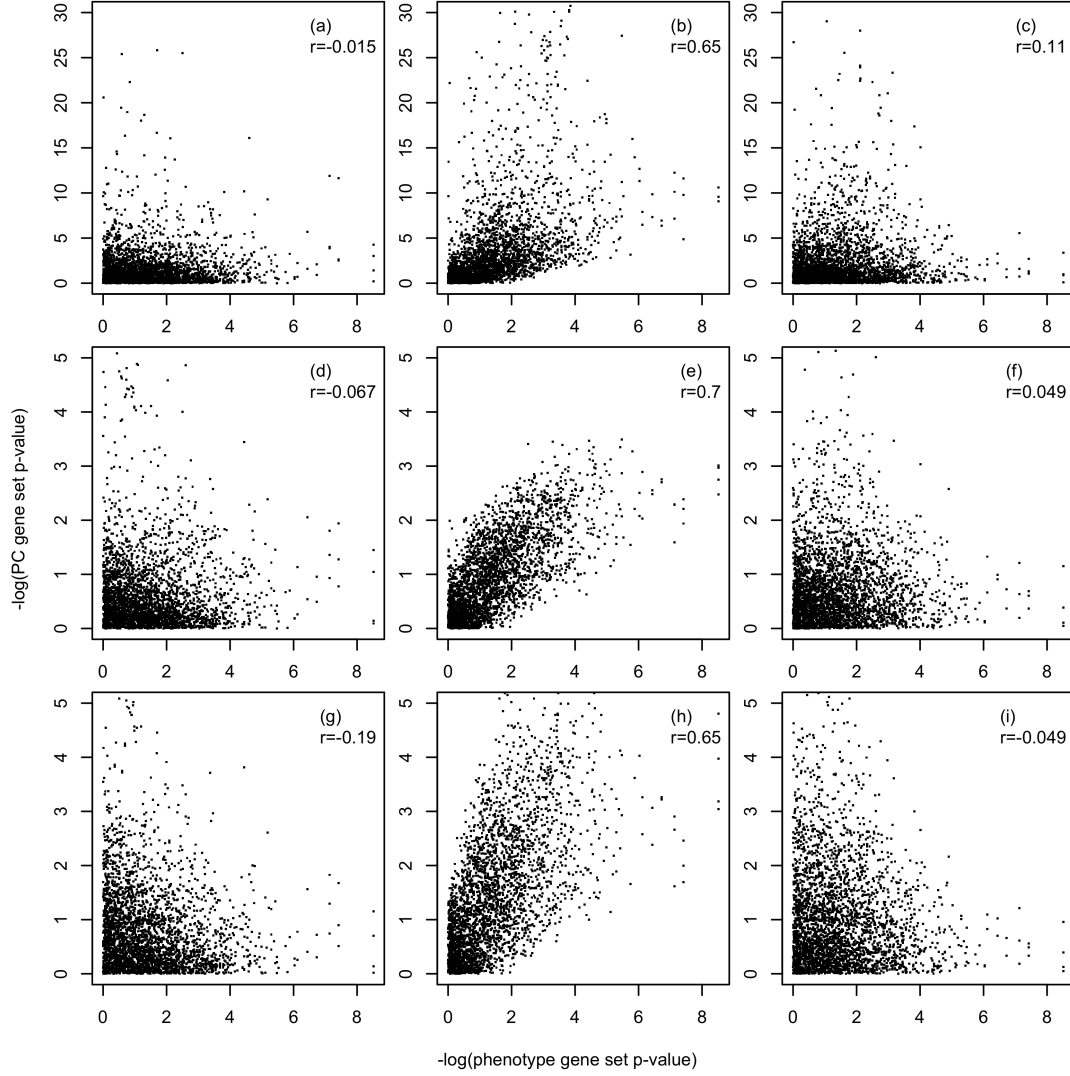


Figure 4: Scatter plots showing the association between phenotype gene set enrichment p-values and PC gene set enrichment p-values for the Armstrong *et al.* (2002) leukemia gene expression data, AML/ALL phenotype, MSigDB C2 v4.0 gene sets and first three PCs. Both phenotype and PC gene set enrichment p-values were computed as outlined in Section 2.9.3. The Spearman correlation coefficient between phenotype and PC gene set enrichment p-values is displayed in each plot. Plots **(a)-(c)** show the association between phenotype and PC gene set enrichment p-values for PCs 1 through 3 with the PC enrichment p-values computed using a two-sided t-test on the standardized mean difference gene set statistic. For plots **(d)-(f)**, the PC gene set enrichment p-values were computed using a correlation-adjusted two-sided t-test and, for plots **(g)-(i)**, the PC gene set enrichment p-values were computed using the permutation distribution of the gene set statistic.

either an unadjusted t-test or permutation test.



## 4 Conclusion

Although principal component analysis is widely used for the dimensional reduction of biomedical data, with applications in visualization, clustering and regression, interpretation of PCA-based models remains challenging. While rotation methods and sparse PCA techniques can generate approximate PCs with few non-zero loadings that support interpretation in terms of individual variables, these approaches will perform poorly on genomic data in which important biological signals are defined by the collective action of groups of functionally related genes. Although gene set testing methods have been widely applied to analyze the association between gene sets and clinical phenotypes, such variable group testing methods have seen little application for testing the association between gene sets and the spectra of genomic data. To address the challenge of PC interpretation for genomic data and support the interpretation of genomic PCs in terms of functional gene sets, we have developed the principal component gene set enrichment (PCGSE) method. PCGSE performs a two-stage competitive gene set test using the correlation between each gene and each PC as the gene-level test statistic with flexible choice of both the gene set test statistic and the method used to compute the null distribution of the gene set statistic. To facilitate use of the PCGSE method by other researchers, an implementation of the technique is available as an R package from CRAN. On both simulated gene sets with simulated data and on curated gene sets with real gene expression data, a computationally efficient version of the PCGSE method based on a correlation-adjusted two-sided, two-sample t-test has been shown to accurately compute the statistical association between gene sets and the PCs of genomic data.

## Acknowledgement

**Funding:** National Institutes of Health R01 grants LM010098, LM011360, EY022300, GM103506 and GM103534.

**Conflict of Interest:** None declared.

## References

- Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.
- Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, **7**(1), 55–65.
- Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, **97**(18), 10101–10106.
- Anderson, T. W. (1963). Asymptotic Theory for Principal Component Analysis. *Annals of Mathematical Statistics*, **34**(1), 122–148.
- Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., and Korsmeyer, S. J. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, **30**(1), 41–7.

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.
- Barry, W. T., Nobel, A. B., and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**(9), 1943–9.
- Barry, W. T., Nobel, A. B., and Wright, F. A. (2008). A statistical framework for testing functional categories in microarray data. *Annals of Applied Statistics*, **2**, 286–315+.
- Bruckskotten, M., Looso, M., Cemić, F., Konzer, A., Hemberger, J., Krüger, M., and Braun, T. (2010). PCA2GO: a new multivariate statistics based method to identify highly expressed GO-Terms. *BMC Bioinformatics*, **11**, 336.
- Chen, X. (2011). Adaptive elastic-net sparse principal component analysis for pathway association testing. *Statistical Applications In Genetics and Molecular Biology*, **10**(1), 48.
- d’Aspremont, A., El Ghaoui, L., Jordan, M. I., and Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, **49**(3), 434–448.
- Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Annals of Applied Statistics*, **1**(1), 107–129.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *ArXiv e-prints*. arXiv:1001.0736.
- Goeman, J. J. and Buehlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**(8), 980–987.
- Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**(1), 93–9.
- Grbovic, M., Dance, C., and Vucetic, S. (2012). Sparse principal component analysis with constraints. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 953–941.
- Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. (2000). ‘gene shaving’ as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **1**(2), RESEARCH0003.
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York, NY, 2nd ed edition.
- Hibbs, M. A., Dirksen, N. C., Li, K., and Troyanskaya, O. G. (2005). Visualization methods for statistical analysis of microarray clusters. *BMC Bioinformatics*, **6**, 115.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**.
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**(1), 1–13.

- Jenatton, R., Obozinski, G., and Bach, F. (2010). Structured sparse principal component analysis. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 366–373.
- Jolliffe, I. (1995). Rotation of Principal Components - Choice of Normalization Constraints. *Journal of Applied Statistics*, **22**(1), 29–35.
- Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer.
- Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**(3), 531–547.
- Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**(1), 27–30.
- Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, **8**(2), e1002375.
- Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003). Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, **13**(4), 703–716.
- Kong, S. W., Pu, W. T., and Park, P. J. (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**(19), 2373–2380.
- Liu, J. and Ye, J. (2010). Fast Overlapping Group Lasso. *ArXiv e-prints*. arXiv:1009.0306.
- Lu, J., Kerns, R. T., Peddada, S. D., and Bushel, P. R. (2011). Principal component analysis-based filtering improves detection for affymetrix gene expression arrays. *Nucleic Acids Research*, **39**(13), e86.
- Ma, S. and Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*, **12**(6, SI), 714–722.
- Ma, S. and Kosorok, M. R. (2009). Identification of differential gene pathways with principal component analysis. *Bioinformatics*, **25**(7), 882–889.
- Moghaddam, B., Weiss, Y., and Avidan, S. (2006). Spectral bounds for sparse pca: Exact and greedy algorithms. *Advances in neural information processing systems*, **18**, 915.
- Obozinski, G., Jacob, L., and Vert, J.-P. (2011). Group Lasso with Overlaps: the Latent Group Lasso approach. *ArXiv e-prints*. arXiv:1110.0413.
- Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLOS Genetics*, **2**(12), e190.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**(6), 559–572.
- Ramsay, J., Berge, J., and Styán, G. (1984). Matrix correlation. *Psychometrika*, **49**, 403–423.
- Roden, J. C., King, B. W., Trout, D., Mortazavi, A., Wold, B. J., and Hart, C. E. (2006). Mining gene expression data by interpreting principal components. *BMC Bioinformatics*, **7**, 194.
- Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, **99**(6), 1015–1034.

- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, **9**(12), 3273–3297.
- Sriperumbudur, B. K., Torres, D. A., and Lanckriet, G. R. G. (2011). A majorization-minimization approach to the sparse generalized eigenvalue problem. *Machine Learning*, **85**(1-2), 3–39.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**(43), 15545–15550.
- Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*, **102**(38), 13544–9.
- Tomfohr, J., Lu, J., and Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.
- Vines, S. (2000). Simple principal components. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **49**(Part 4), 441–451.
- Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**(3), 515–534.
- Wu, D. and Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, **40**(17), e133.
- Yanai, H. (1980). A proposition of generalized method for forward selection of variables. *Behaviormetrika*, **7**(7), 95–107.
- Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**(9), 763–774.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **68**(Part 1), 49–67.
- Zhou, Y.-H., Barry, W. T., and Wright, F. A. (2013). Empirical pathway analysis, without permutation. *Biostatistics*, **14**(3), 573–85.
- Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.